MÓDULO 6:

ANÁLISIS MULTIVARIADO

PROFESOR: LUIS E. NIETO BARAJAS

EMAIL: lnieto@itam.mx

URL: http://allman.rhon.itam.mx/~lnieto

Diplomado en Estadística Aplicada

Módulo 6: Análisis Multivariado

- OBJETIVO: Proporcionar al alumno los aspectos básicos de la teoría y de la aplicación con computadora de las principales técnicas del análisis estadístico de varias variables (multivariado).

- PLAN DE ESTUDIOS:

  1. Introducción.

  2. Análisis exploratorio multivariado.

  3. Análisis de componentes principales.

  4. La distribución normal multivariada.

  5. Análisis de factores.

  6. Análisis de cúmulos.

  7. Escalamiento multidimensional.

  8. Análisis de varias variables categóricas.

  9. Solución de problemas prácticos.

- REFERENCIA BÁSICA:

✓ Johnson, D. E. (2000). Métodos multivariados aplicados. ITP International Thomson Editores: México.

- REFERENCIAS ADICIONALES:

- Kachigan, S. K. (1991). Multivariate statistical analysis. Radius Press.
- Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. (1998). Multivariate data analysis. Prentice Hall College Division.
- Johnson, R. A. & Wichern, D. W. (2002). Applied multivariate statistical analysis. Prentice Hall: London.

- PAQUETES ESTADÍSTICOS: En el curso habrá un paquete estadístico básico, en el cual se ejemplificarán las técnicas presentadas. Este paquete básico no es exclusivo, si el alumno así lo desea, puede auxiliarse de cualquier otro paquete estadístico.
  - ✓ *Paquete básico*: Splus
  - ▫ *Paquetes auxiliares*: SPSS, Statgraphics, Minitab, Matlab

- EVALUACIÓN: El alumno debe de buscar una base de datos (multivariada) a la cual le puede ir aplicando las distintas técnicas multivariadas que se vayan desarrollando a lo largo del curso. Al finalizar el módulo, el alumno deberá entregar un trabajo conteniendo los siguientes puntos:

  1) Descripción de la base de datos.
  2) Análisis de cada una de las técnicas multivariadas vistas en clase.
  3) Conclusiones sobre los análisis realizados en el contexto de los datos.
  4) Fuente de los datos y bibliografía usada.

1. Introducción

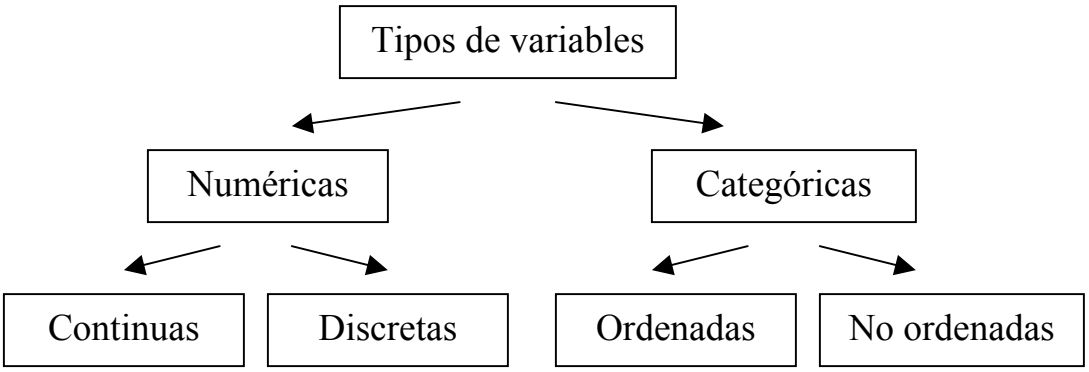
- Los datos multivariados surgen en distintas áreas o ramas de la ciencia. Ejemplos:
  1) Investigación de mercados: Identificar características de los individuos para determinar qué tipo de personas compran determinado producto.
  2) Agricultura: Resistencia de determinado tipo de cosechas a daños por plagas y sequías.
  3) Psicología: Relación entre el comportamiento de adolescentes y actitudes de los padres.

- ¿En qué situaciones surgen los datos multivariados?
  Cuando a un mismo individuo se le mide más de una característica de interés.

- Un *individuo* puede ser un objeto o concepto que se puede medir. Más generalmente, los individuos son llamados *unidades experimentales*.
  Ejemplos de objetos: personas, animales, terrenos, compañías, países, etc.
  Ejemplos de conceptos: amor, amistad, noviazgo, etc.

  - Características de los individuos: Los individuos deben de ser *independientes* entre sí.

- Una *variable* es una característica o atributo que se le mide a un individuo.



- OBJETIVOS de los métodos multivariados:

1) *Simplificación*: Los métodos multivariados son un conjunto de técnicas que permiten al investigador interpretar y visualizar conjuntos grandes de datos (tanto en individuos como en variables).

2) *Relación*: Encontrar relaciones entre variables, entre individuos y entre ambos.

2.1) Relación entre variables: Existe relación entre variables cuando las variables miden una característica común. Ejemplo: Suponga que se realizan exámenes de lectura, ortografía, aritmética y álgebra a estudiantes de 6^{o} de primaria. Si cada uno de los estudiantes obtiene calificaciones altas, regulares o bajas en los cuatro exámenes, entonces los exámenes estarían relacionados entre sí. En este caso, la característica común que estos exámenes pueden estar midiendo podría ser la "inteligencia global".

2.2) Relación entre individuos: Existe relación entre individuos si alguno de ellos son semejantes entre sí. Ejemplo: Suponga que se evalúan cereales (para el desayudo) respecto a su contenido nutricional y se miden, por ejemplo, los gramos de grasa, proteínas, carbohidratos y sodio a cada uno de ellos. Se podría esperar que los cereales de fibra estén relacionados entre sí, o que los cereales

endulzados tengan cierta relación entre sí, además se podría esperar que ambos grupos fueran diferentes de uno a otro.

- *Uso* de los métodos multivariados: Minerías de datos (*data mining*).

- Los métodos multivariados son realmente un conjunto de técnicas que en su gran mayoría tienen un carácter exploratorio y no tanto inferencial.

- CLASIFICACIÓN de los métodos multivariados:
  1) *Dirigidas o motivadas por las variables*: se enfocan en las relaciones entre variables. Ejemplos: matrices de correlación, análisis de componentes principales, análisis de factores, análisis de regresión y análisis de correlación canónica.
  2) *Dirigidas o motivadas por los individuos*: se enfocan en las relaciones entre individuos. Ejemplos: análisis discriminante, análisis de cúmulos y análisis multivariado de varianza.

- EJEMPLOS de datos multivariados.

- *Ejemplo 1*. (Johnson, 2000). Características de candidatos a ingresar a la policía.

  Variables (medidas en centímetros).

  EST: Estatura

  ESTSEN: Estatura sentados

  BRAZO: Longitud del brazo

  ANTEB: Longitud del antebrazo

  MANO: Ancho de la mano

MUSLO: Longitud del muslo

PIERNA: Longitud de la parte inferior de la pierna

PIE: Longitud del pie

Variables adicionales:

BRACH: Razón de la longitud del antebrazo y de la del brazo × 100

TIBIO: Razón de la parte inferior de la pierna y la del muslo × 100

- *Ejemplo 2*. (Johnson, 2000). Consumo de caucho y otras variables desde 1948 hasta 1963.

  Variables.

  CTC: Consumo total de caucho

  CCN: Consumo de caucho para neumáticos

  PA: Producción de automóviles

  PNB: Producto nacional bruto

  IPD: Ingreso personal disponible

  CCM: Consumo de combustible por motor

- *Ejemplo 3*. (SIMM90, CONAPO). Sistema automatizado de información sobre la marginación en México 1990.

  Variables.

  CLAVE: Clave geoestadística

  NOMBRE: Nombre

  POB: Población total

  SUPERF: Superficie

  DENSP: Densidad

  ANALF: Porcentaje de población mayor de 15 años analfabeta

  S/PRI: Porcentaje de población mayor de 15 años sin primaria completa

S/EXC: Porcentaje de ocupantes en viviendas sin drenaje ni excusado

S/ELE: Porcentaje de ocupantes en viviendas sin energía eléctrica

S/AGU: Porcentaje de ocupantes en viviendas sin agua entubada

HACIN: Porcentaje de viviendas con hacinamiento

PISOT: Porcentaje de ocupantes en viviendas con piso de tierra

L5000: Porcentaje de población en localidades con menos de 5,000 habitantes

INGRE: Porcentaje de población ocupada con ingreso menor de 2 salarios mínimos

INDICE: Indice de marginación

GRADO: Grado de marginación

- NOTACIÓN de matrices y vectores:

  p = número de variables

  n = número de individuos

  X_{ij} = j-ésima variable del i-ésimo individuo

  x_{ij} = valor observado de la j-ésima variable del i-ésimo individuo

  i=1,...,n y j=1,....,p

- *Matriz de datos*:

$$x = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

  x_{ij} = elemento en el i-ésimo renglón y j-ésima columna

  Renglones = individuos

  Columnas = variables

- *Vectores de datos*:

  Los renglones de la matriz de datos se pueden expresar como vectores de la siguiente forma: El i-ésimo renglón de X se escribe como

$$x_i' = (x_{i1}, x_{i2}, ..., x_{ip})$$

  Nota: Todos los vectores son vectores columna, i.e.,

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

- ESPERANZAS y VARIANZAS de vectores aleatorios

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

- *Media*:

$$\mu = E(X) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

  μ es un vector de medias de dimensión $p \times 1$.

- *Varianzas-Covarianzas*:

$$\Sigma = Cov(X) = Var(X) = E\left\{(X - \mu)(X - \mu)'\right\}$$

Escribiendo el vector completo,

$$\Sigma = \mathrm{E}\left\{ \begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{pmatrix} (X_1 - \mu_1, X_2 - \mu_2, ..., X_p - \mu_p) \right\}$$

$$= \mathrm{E}\begin{pmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \cdots & (X_1 - \mu_1)(X_p - \mu_p) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \cdots & (X_2 - \mu_2)(X_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p - \mu_p)(X_1 - \mu_1) & (X_p - \mu_p)(X_2 - \mu_2) & \cdots & (X_p - \mu_p)^2 \end{pmatrix}$$

Finalmente, los elementos de Σ se denotan como:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

donde, $\sigma_{jj} = \mathrm{Cov}(X_j, X_j) = \mathrm{Var}(X_j) = \mathrm{E}\{(X_j - \mu_j)^2\}$, para j=1,2,...,p, y

$\sigma_{kj} = \mathrm{Cov}(X_k, X_j) = \mathrm{E}\{(X_k - \mu_k)(X_j - \mu_j)\}$, para k≠j=1,2,...,p

Σ es una matriz de varianzas y covarianzas dimensión n×p.

- *Correlaciones*:

$$\mathrm{P} = \mathrm{Corr}(X) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

donde, $\rho_{kj} = \mathrm{Corr}(X_k, X_j) = \dfrac{\sigma_{kj}}{\sqrt{\sigma_{kk}}\sqrt{\sigma_{jj}}}$, para k≠j=1,2,...,p

Cometarios:

1) El coeficiente de correlación ρ_{kj} es una medida de la *relación lineal* entre las variables X_k y X_j.

2) $-1 \leq \rho_{kj} \leq 1$

3) Si X_k y X_j son v.a. independientes $\Rightarrow \rho_{kj} = 0$.

4) $\rho_{kj} = 0 \Rightarrow$ Independencia entre X_k y X_j únicamente en el caso Normal.

5) Para apreciar la relación (en general) entre dos variables es recomendable, además de calcular en coeficiente de correlación, hacer una gráfica de dispersión de ellas.

2. Análisis exploratorio multivariado

2.1. Estadísticas multivariadas descriptivas

- Las estadísticas descriptivas (multivariadas), como su nombre lo indica, sirven para describir el comportamiento de un conjunto de datos.

- Formalmente, un *conjunto de datos* es una realización de una muestra aleatoria $X_1, X_2, ..., X_n$ de una distribución multivariada. Es decir, para i=1,...,n,

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{pmatrix}.$$

  En otras palabras, cada X_i es una variable aleatoria multivariada de dimensión *p*.

  - Por lo tanto, un conjunto de datos esta formado por *n* realizaciones de *p* variables aleatorias.

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}.$$

- MEDIA MUESTRAL:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i ,$$

que en realidad, escribiendo el vector completo, se puede expresar como:

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_p \end{pmatrix} = \frac{1}{n}\left\{ \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1p} \end{pmatrix} + \cdots + \begin{pmatrix} X_{n1} \\ X_{n2} \\ \vdots \\ X_{np} \end{pmatrix} \right\}.$$

Esto implica que, para j=1,...,p

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij} .$$

- *Propiedades*: $E(\hat{\mu}) = \mu$.

❖ Splus: mean

➢ VARIANZA MUESTRAL:

$$\hat{\Sigma} = \frac{1}{n-1}\left\{ \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})' \right\},$$

cuyos elementos se denotan como:

$$\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1p} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{p1} & \hat{\sigma}_{p2} & \cdots & \hat{\sigma}_{pp} \end{pmatrix}$$

donde, $\hat{\sigma}_{jj} = \frac{1}{n-1}\sum_{i=1}^{n} (X_{ij} - \hat{\mu}_j)^2$, para j=1,2,...,p, y

$$\hat{\sigma}_{kj} = \frac{1}{n-1}\sum_{i=1}^{n} (X_{ik} - \hat{\mu}_k)(X_{ij} - \hat{\mu}_j), \text{ para } k \neq j = 1,2,...,p.$$

- *Propiedades*: $E(\hat{\Sigma}) = \Sigma$.

❖ Splus: var

- CORRELACIÓN MUESTRAL:

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

donde, $r_{kj} = \dfrac{\hat{\sigma}_{kj}}{\sqrt{\hat{\sigma}_{kk}}\sqrt{\hat{\sigma}_{jj}}}$, para $k \neq j = 1,2,...,p$.

- *Propiedades*:

  1) $-1 \leq r_{kj} \leq 1$

  2) $E(R) \neq P$.

❖ Splus: cor

- CUARTILES MUESTRALES: Estas estadísticas de orden se obtienen como en el caso univariado para cada una de las variables.

❖ Splus: summary

2.2. Análisis gráfico de datos multivariados

- DIAGRAMAS DE DISPERSIÓN (bidimensional)

  Este tipo de diagrama consiste en graficar simultáneamente en dos dimensiones diagramas de dispersión entre todas las posibles parejas de variables.

❖ Splus: plot, pairs

- DIAGRAMAS DE DISPERSIÓN (tridimensional)

  Este tipo de diagrama consiste en graficar en tres dimensiones tres variables simultáneamente.

  - Splus: Graph > 3D Plot > 3D Scatter Plot

- DIAGRAMA DE BURBUJAS (tridimensional)

  Este tipo de diagrama consiste en graficar en dos dimensiones tres variables en forma de burbujas de la siguiente manera: El eje de las X's corresponde a una de las variables, el eje de las Y's corresponde a otra de las variables, y la tercer variable quedará representada por el tamaño de la burbuja.

  - Splus: symbols

- CARAS DE CHERNOFF (multidimensional)

  Este tipo de diagrama consiste en graficar un conjunto multivariado de variables en forma de caras, asociando características faciales diferentes a variables diferentes. Por ejemplo, una variable se podría asociar con el ancho vertical del ojo, la segunda con el ancho horizontal, la tercera con el tamaño del iris, y las otras se podrían asociar con el espaciamiento de los ojos, la altura de los ojos, la longitud de la nariz, en ancho de la nariz, la longitud de las cejas, el ancho de las cejas. La inclinación de las cejas, el ancho de las orejas, la longitud de las orejas, la abertura de la boca, la sonrisa, etc.

  - Estos diagramas son útiles para detectar datos extremos (*outliers*).
  - Splus: faces

- DIAGRAMA DE ESTRELLAS (multidimensional)

  Este tipo de diagrama se aplica cuando todas las variables toman valores positivos y consisten en graficar rayos o ejes que parten de un punto central. La longitud del rayo corresponde al valor de la variable y se tiene un rayo para cada variable. Por ejemplo, vectores de datos con 5 variables requerirán 5 rayos separados entre sí por un ángulo de 72 grados.

  La primera variable generalmente corresponde con el rayo que apunta hacia el norte y las otras variables se representan sobre los otros rayos en el orden del sentido del movimiento de las manecillas del reloj.

❖ Splus: stars

- DIAGRAMA DE ANDREWS (multidimensional)

  Este tipo de diagrama consiste en representar a la observación i-ésima de un vector aleatorio p-variado $x_i' = (x_{i1}, x_{i2}, ..., x_{ip})$ de la siguiente forma:

$$f_i(t) = \frac{x_{i1}}{\sqrt{2}} + x_{i2}\text{sen}(t) + x_{i3}\cos(t) + x_{i4}\text{sen}(2t) + x_{i5}\cos(2t) + \cdots$$

  para $-\pi < t < \pi$. De esta forma, las observaciones para el individuo i dan lugar a una única función $f_i(t)$. El diagrama de Andrews se construye graficando las funciones $f_1(t)$, $f_2(t)$,... $f_n(t)$ para $-\pi < t < \pi$.

- Estos diagramas son útiles para encontrar agrupamientos en los datos. También son útiles para localizar datos extremos.
- Es recomendable que las variables estén medidas en unidades semejantes (estandarización).
- El orden de las variables afecta la interpretación.

3. Análisis de componentes principales.

3.1. Breve repaso de matrices

- Sea Σ una matriz cuadrada de p×p tal que

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}.$$

- Se dice que una matriz es *simétrica* si $\sigma_{jk} = \sigma_{kj}$ para todo j,k=1,2,...,p.

  Las matrices de varianzas-covarianzas siempre son simétricas.

- *Traza* de una matriz: $\text{tr}(\Sigma) = \sum_{j=1}^{p} \sigma_{jj}$.

- *Determinante* de una matriz (cuadrada): $\det(\Sigma) = |\Sigma| = \sum_{j=1}^{p} \sigma_{1j}\Sigma_{1j}$, en donde $\Sigma_{1j} = (-1)^{1+j}\left|\Sigma^{1j}\right|$ y Σ^{1j} es la matriz obtenida a partir de Σ al eliminar su primer renglón y su j-ésima columna.
  - El determinante de una matriz de 1×1 es igual al valor del único elemento. Ej: Si $\Sigma = (\sigma_{11})$ entonces $|\Sigma| = \sigma_{11}$.
  - El de terminante de una matriz de 2×2 se calcula como:

    Si $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ entonces $|\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}$.

- *Ejemplo numérico*:

  Sea $\Sigma = \begin{pmatrix} 6 & 2 \\ 2 & 3 \end{pmatrix}$.

  Entonces, tr(Σ)=6+3=9, y $|\Sigma| = 6(-1)^{1+1}|3| + 2(-1)^{1+2}|2| = 18 - 4 = 14$.

❖ Splus: det

➢ *Eigenvalores* y *eigenvectores*: Los eigenvalores (o valores característicos) y los eigenvectores (o vectores característicos) son valores y vectores que caracterizan una matriz (cuadrada) y satisfacen

$$\Sigma w = \lambda w, \tag{3.1}$$

donde λ es un eigenvalor y w es un eigenvector.

- Los eigenvalores se obtienen como solución a la ecuación:

$$|\Sigma - \lambda I| = 0,$$

donde I es la matriz identidad. Esta expresión toma la forma de una ecuación polinomial en λ de grado p:

$$c_1\lambda^p + c_2\lambda^{p-1} + \cdots + c_p\lambda + c_{p+1} = 0.$$

Las raíces de esta ecuación son los eigenvalores de Σ. En general, $\lambda' = (\lambda_1, \lambda_2, ..., \lambda_p)$.

- Si Σ es una matriz simétrica, sus eigenvalores son número reales y por lo tanto se pueden ordenar de forma descendente $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.
- Para cada eigenvalor λ_j, existe un eigenvector w_j que satisface la ecuación (3.1).

- *Propiedades*:

$$\text{tr}(\Sigma) = \sum_{j=1}^{p} \lambda_j ,$$

$$|\Sigma| = \prod_{j=1}^{p} \lambda_j .$$

- *Ejemplo numérico*:

  Sea $\Sigma = \begin{pmatrix} 6 & 2 \\ 2 & 3 \end{pmatrix}$.

  Los eigenvalores de Σ deben satisfacer $|\Sigma - \lambda I| = 0$, es decir,

$$\left| \begin{pmatrix} 6-\lambda & 2 \\ 2 & 3-\lambda \end{pmatrix} \right| = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

  Esto implica que $(6-\lambda)(3-\lambda)-4=0$, por lo que $\lambda^2 - 9\lambda + 14 = 0$. Resolviendo la ecuación obtenemos que $\lambda_1=7$ y $\lambda_2=2$.

  Para calcular el eigenvector correspondiente a $\lambda_1=7$ hacemos, $\Sigma w_1 = \lambda_1 w_1$, es decir,

$$\begin{pmatrix} 6 & 2 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} w_{11} \\ w_{21} \end{pmatrix} = 7 \begin{pmatrix} w_{11} \\ w_{21} \end{pmatrix} \Rightarrow \begin{matrix} 6w_{11} + 2w_{21} = 7w_{11} \\ 2w_{11} + 3w_{21} = 7w_{21} \end{matrix} \Rightarrow w_{11} = 2w_{21}.$$

  Existen muchos vectores que satisfacen la condición $w_{11} = 2w_{21}$, pero el único vector normalizado $\left(w_1' w_1 = 1\right)$ es: $w_1' = \left(2/\sqrt{5}, 1/\sqrt{5}\right)$.

  Similarmente, resolviendo $\Sigma w_2 = \lambda_2 w_2$ para $\lambda_2=2$ se puede demostrar que $w_2' = \left(1/\sqrt{5}, -2/\sqrt{5}\right)$.

- Una matriz es *definida positiva* si todos sus eigenvalores son positivos.

- Una matriz es *semi-definida positiva* si todos sus eigenvalores son no negativos.

- NOTA: Las matrices de varianzas-covarianzas y de correlaciones tanto poblacionales como muestrales son semidefinidas positivas.

3.2. Componentes principales

- El análisis de componentes principales es un procedimiento matemático que transforma un conjunto de variables posiblemente correlacionadas en un conjunto menor de variables no correlacionadas llamadas *componentes principales*.

- Dadas n observaciones de p variables, el objetivo del análisis de componentes principales es determinar r nuevas variables no correlacionadas llamadas componentes principales que representen la mayor variabilidad posible de las variables originales.

- El uso de esta técnica es principalmente exploratoria y en general como un paso intermedio para análisis posteriores.

- Los *objetivos* principales son:
  1) Reducir la dimensionalidad de un conjunto de datos,
  2) Interpretar un conjunto de datos.

- CARACTERÍSTICAS: Las nuevas variables (componentes principales) son creadas de tal manera que:

1) No estén correlacionadas.

2) La 1ª componente principal explique la mayor variabilidad posible de los datos.

3) Cada componente subsecuente explique la mayor variabilidad posible restante no explicada por las componentes anteriores.

- Formalmente, sea $X' = (X_1, X_2, ..., X_p)$ un vector aleatorio de *p* variables con matriz de varianzas-covarianzas Σ con eigenvalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Sean $Y' = (Y_1, Y_2, ..., Y_p)$ nuevas variables formadas como combinaciones lineales de las X_i's, i.e.,

$$\begin{aligned} Y_1 &= a_1'X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 &= a_2'X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ Y_p &= a_p'X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned}$$

Las componentes principales son aquellas combinaciones lineales $Y_1, Y_2, ..., Y_p$ no correlacionadas, cuyas varianzas son tan grandes como sea posible.

- COMPONENTES:

1ª componente principal: $Y_1 = a_1'X$, donde a_1 maximiza $\text{Var}(a_1'X)$ sujeto a

$$a_1'a_1 = 1$$

2ª componente principal: $Y_2 = a_2'X$, donde a_2 maximiza $\text{Var}(a_2'X)$ sujeto a

$$a_2'a_2 = 1 \text{ y } \text{Cov}(a_1'X, a_2'X) = 0$$

kª componente principal: $Y_k = a_k'X$, donde a_k maximiza $\text{Var}(a_k'X)$ sujeto a

$$a_k'a_k = 1 \text{ y } \text{Cov}(a_k'X, a_j'X) = 0 \text{ para } j<k$$

- Se puede demostrar que el máximo de la varianza de $a_1'X$ entre todos los vectores a_1 que satisfacen $a_1'a_1 = 1$ es igual a λ_1 y por lo tanto, a_1 es el eigenvector de Σ correspondiente al eigenvalor λ_1.
- También, se puede demostrar que el valor máximo de la varianza de $a_2'X$ entre todas las combinaciones lineales que satisfacen $a_2'a_2 = 1$ y que no están correlacionadas con Y_1 es igual a λ_2. Por lo tanto, a_2 es el eigenvector de Σ correspondiente al eigenvalor λ_2.
- En general, se puede demostrar que el valor máximo de la varianza de $a_k'X$ entre todas las combinaciones lineales que satisfacen $a_k'a_k = 1$ y que no están correlacionadas con $Y_1, Y_2, ..., Y_{k-1}$ es igual a λ_k. Por lo tanto, a_k es el eigenvector de Σ correspondiente al eigenvalor λ_k.

➢ INTERPRETACIÓN de λ_k:

Recuerde que $\text{tr}(\Sigma) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp}$ es una medida de la variabilidad total de las variables originales. Por otro lado, $\text{Var}(Y_k) = \text{Var}(a_k'X) = \lambda_k$, k=1,...,p. Por lo tanto, la variabilidad total de las variables componentes principales $\text{tr}(\Sigma) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$ es igual a la variabilidad total de las variables originales.

- $\begin{pmatrix} \text{Proporción de la variabilidad} \\ \text{total explicada por la k - ésima} \\ \text{componente principal} \end{pmatrix} = \dfrac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}, \quad k=1,2,...,p$

- INTERPRETACIÓN del vector de pesos $a_k' = (a_{k1}, a_{k2}, ..., a_{kp})$:

  Los elementos a_{kj} del eigenvector a_k son llamados *pesos* y miden la importancia de la j-ésima variable en el k-ésimo componente principal.

- ¿*Cuántos* componentes principales son suficientes?

  El número de componentes principales que de alguna manera pudieran reemplazar a las variables originales, sin mucha pérdida de información, depende del problema en particular. En general, se desea que el porcentaje de la variabilidad explicada por los *r* primeros componentes sea de al menos el 80%.

  - Una forma alternativa de decidir el número de componentes significativos es graficando k vs. λ_k. Cuando los puntos de la gráfica tienden a nivelarse, estos eigenvalores suelen estar suficientemente cercanos a cero como para que puedan ignorarse.

- NOTA: Si no se tiene la matriz de varianzas-covarianzas poblacional Σ, se realiza todo el análisis anterior sobre la matriz de varianzas-covarianzas muestral $\hat{\Sigma}$. En este caso, los componentes obtenidos serían estimaciones de los componentes poblacionales.

- *Valores* o marcadores (*scores*) de los componentes principales: Para poder visualizar las componentes principales es necesario calcular el valor de cada componente para cada individuo en un conjuntode datos.

  Sea x_i el vector de variables medidas para cada individuo. Entonces el valor de la k-ésima componente principal para el i-ésimo individuo es

$$y_{ik} = a_k' x_i, \text{ para } i=1,...,n \text{ y } k=1,...,p.$$

3.3. Componentes principales sobre variables estandarizadas

- Si la escala en que están medidas las variables no es uniforme (similar), es recomendable realizar un análisis de componentes principales sobre las variables estandárizadas, i.e.,

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}},\ Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}},\ \ldots,\ Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

  En notación matricial, $Z = \Sigma^{-1/2}(X - \mu)$.

  - Propiedades: $E(Z) = 0$ y $Var(Z) = Cov(Z) = P$, donde P es la matriz de correlaciones de los datos originales X.

- Los componentes principales $Y^{*'} = (Y_1^*, Y_2^*, \ldots, Y_p^*)$ del conjunto de variables estandarizadas $Z' = (Z_1, Z_2, \ldots, Z_p)$ se obtienen de los eigenvectores de la matriz de correlación P de X.

- Vectores de correlaciones de componentes:

  Si λ_k^* y a_k^* son los eigenvalores y eigenvectores de la matriz P, las cantidades $c_k = \lambda_k^{*\,1/2} a_k^*$ dan las correlaciones entre las variables originales y la k-ésima componente principal, i.e.,

$$Corr(Y_k^*, X_j) = c_{kj},\ \text{para } j,k=1,\ldots,p.$$

- NOTA: Los componentes principales obtenidos a partir de la matriz Σ son, en general, diferentes a los obtenidos de la matriz P.

❖ Splus: princomp, print.summary.princomp

4. La distribución normal multivariada.

4.1. Introducción y definiciones.

- La mayoría de los métodos multivariados tradicionales cuando son usados para realizar inferencias, mas que para un carácter exploratorio, suponen que los vectores de datos son muestras de v.a. normales multivariadas.

- Un vector aleatorio X es normal multivariado si su distribución conjunta es normal multivariada.

- Existen varias DEFINICIONES equivalentes de una distribución normal multivariada:

  - *"Definición 1" (Simple)*: Se dice que un vector aleatorio $X' = (X_1, X_2, ..., X_p)$ tiene una distribución normal multivariada si

$$a'X = (a_1, a_2, ..., a_p)\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = \sum_{j=1}^{p} a_j X_j$$

    tiene una distribución normal univariada para todos los posibles valores del vector a.

  - *Definición 2 (Formal)*: Se dice que un vector aleatorio $X' = (X_1, X_2, ..., X_p)$ tiene una distribución normal multivariada con vector de medias μ y matriz de varianzas-covarianzas Σ, si su función de densidad está dada por

$$f_X(x;\mu,\Sigma)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}, \text{ para } x\in\Re^p$$

- Notación: $X \sim N_p(\mu, \Sigma)$

➢ PROPIEDADES de la distribución normal multivariada:

Si $X \sim N(\mu, \Sigma)$, es decir, el vector $X'=(X_1,X_2,...,X_p)$ tiene una distribución normal multivariada, entonces

1) $E(X) = \mu$ y $Var(X) = \Sigma$.

2) Cada X_j, para j=1,...,p, tiene un distribución normal univariada. Es decir, $X_j \sim N(\mu_j, \sigma_{jj})$ y por lo tanto, $E(X_j) = \mu_j$ y $Var(X_j) = \sigma_{jj}$.

3) Si $\sigma_{jk}=0$ $(\rho_{jk}=0)$ para $j\neq k=1,...,p$ entonces $X_1,X_2,...,X_p$ son v.a. independientes.

- *Nota*: Si cada X_j, j=1,..,p tiene una distribución normal univariada, no necesariamente el vector $X'=(X_1,X_2,...,X_p)$ tendrá una distribución normal multivariada. En general sí se cumple, pero existen algunos casos atípicos en donde no.

4.2. Distribución normal bivariada

➢ Un caso particular de la distribución normal multivariada es cuando el número de variables p=2. En este caso, si $X'=(X_1,X_2) \sim N_2(\mu, \Sigma)$ se dice que X tiene una distribución normal multivariada de dimensión 2 o que X tiene una distribución normal bivariada, donde

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ y } \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

- Recuerda que $\rho_{12} = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}$.

➢ La distribución normal bivariada es de importancia porque es posible visualizar su comportamiento en una gráfica en tres dimensiones.

➢ Características de la función de densidad normal bivariada.

1) Tiene forma acampanada,

2) Las curvas de nivel forman círculos (si $\sigma_{11}=\sigma_{22}$, $\rho_{12}=0$), o elipses.

❖ Splus: dmvnorm, pmvnorm, rmvnorm.

4.3. Inferencia estadística

➢ El problema de *inferencia estadística* consiste en *aproximar* el valor de ciertas características poblacionales (llamadas *parámetros*) por medio de resúmenes (llamados *estadísticas*) generados a partir de la información contenida en una muestra obtenida de la población.

➢ ESTIMACIÓN PUNTUAL: El problema de estimación puntual consiste en proporcionar un valor puntual que aproxime al parámetro de interés. Los métodos clásicos de estimación puntual de parámetros son: método de momentos y método de máxima verosimilitud.

- De los dos métodos antes mencionados, el que produce estimadores con mejores propiedades (insesgamiento, eficiencia, consistencia, etc.), es el método de máxima verosimilitud.

- El *método de máxima verosimilitud* consiste en encontrar el valor de los parámetros que hacen que la muestra observada tenga probabilidad máxima de haberse observado.

- Los estimadores puntuales para el vector de medias μ, la matriz de varianzas-covarianzas Σ y la matriz de correlaciones P de una distribución normal multivariada son la media muestral $\hat{\mu}$, la varianza muestral $\hat{\Sigma}$ y la correlación muestral R, cuyas expresiones son:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{ó} \quad \hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_p \end{pmatrix} = \frac{1}{n}\left\{ \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1p} \end{pmatrix} + \cdots + \begin{pmatrix} X_{n1} \\ X_{n2} \\ \vdots \\ X_{np} \end{pmatrix} \right\},$$

$$\hat{\Sigma} = \frac{1}{n-1}\left\{ \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})' \right\} \quad \text{ó} \quad \hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1p} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{p1} & \hat{\sigma}_{p2} & \cdots & \hat{\sigma}_{pp} \end{pmatrix},$$

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix},$$

donde, $\hat{\sigma}_{jj} = \frac{1}{n-1}\sum_{i=1}^{n} (X_{ij} - \hat{\mu}_j)^2$, para j=1,2,...,p,

$$\hat{\sigma}_{kj} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ik} - \hat{\mu}_k)(X_{ij} - \hat{\mu}_j), \text{ para } k \neq j = 1,2,...,p, \text{ y}$$

$$r_{kj} = \frac{\hat{\sigma}_{kj}}{\sqrt{\hat{\sigma}_{kk}}\sqrt{\hat{\sigma}_{jj}}}, \text{ para } k \neq j = 1,2,...,p.$$

- *Nota*: El estimador $\hat{\mu}$ es el EMV de μ.

  El estimador $\hat{\Sigma}$ no es el EMV Σ, sino $\frac{n-1}{n}\hat{\Sigma}$.

- *Propiedades*: $E(\hat{\mu}) = \mu$, $E(\hat{\Sigma}) = \Sigma$ y $E(R) \neq P$.

❖ Splus: mean, var, cor.

➢ PRUEBAS DE HIPÓTESIS: El problema de contraste de hipótesis en estadística consiste en decidir cuál de dos hipótesis es correcta. La decisión se toma de acuerdo con la información de la muestra.

➢ La prueba de hipótesis de mayor importancia en datos multivariados es probar si la correlación entre dos variables es significativamente distinta de cero.

➢ *Prueba de hipótesis* para ρ_{jk}: Formalmente, se quiere probar

$$H_0 : \rho_{jk} = 0 \quad \text{vs.} \quad H_1 : \rho_{jk} \neq 0$$

La estadística de prueba es:

$$T = \frac{r_{jk}\sqrt{n-2}}{\sqrt{1-r_{jk}^2}},$$

y la región de rechazo es:

$$\left\{t : |t| > t_{(n-2)}^{\alpha/2}\right\},$$

donde $t_{(n-2)}^{\alpha/2}$ es el punto de una distribución T-Student con (n-2) grados de libertad que acumula $\alpha/2$ de probabilidad a la derecha.

- Splus: cor.test

- INTERVALOS DE CONFIANZA: El calcular un intervalo de confianza es un problema de estimación por intervalo, en donde lo que se proporciona es un conjunto de valores áltamente posibles como aproximaciones al parámetro.

- Al igual que en el caso de pruebas de hipótesis, el intervalo de confianza de mayor interés es el de la correlación entre dos variables.

- *Intervalos de confianza* para ρ_{jk}: Existen varias propuestas, pero una de ellas es la propuesta por Fisher. El intervalo de confianza en este caso sería,

$$\tanh\left\{\tanh^{-1}\left(r_{jk}\right)-\frac{z_{\alpha/2}}{\sqrt{n-3}}\right\}<\rho_{jk}<\tanh\left\{\tanh^{-1}\left(r_{jk}\right)+\frac{z_{\alpha/2}}{\sqrt{n-3}}\right\},$$

donde $z_{\alpha/2}$ es el punto de una distribución normal estándar que acumula $\alpha/2$ de probabilidad a la derecha.

- *Uso de correlaciones* para agrupar variables. Es posible que cuando se tiene un conjunto grande de variables, exista cierta relación entre algunas de las variables. El coeficiente de correlación entre parejas de variables permite agrupar variables de tal manera que variables en el mismo grupo tengan correlaciones altas y variables en grupos diferentes tengan correlaciones bajas.